Analysis of RNA-Seq Data with Partek® Genomics Suite™ 6.6

Overview

RNA-Seq is an emerging technology that uses high-throughput sequencing to generate information about a sample's RNA content. Partek® Genomics Suite™ (PGS) offers convenient visualization and analysis of the high volumes of data generated by RNA-Seq experiments.

This tutorial will illustrate how to:

- Import large next-gen data sets
- Add attribute data to your files
- Visualize large next-gen data sets
- Get read counts for each of the transcripts in a database
- Find transcripts that are differentially expressed among phenotypes
- Find genes that are alternatively spliced among phenotypes
- Set up a basic analysis of variance (ANOVA) model
- Detect nucleotide variations across samples or comparing to reference genome
- Find nonannotated regions and map it to the genome

Note: the workflow described below is enabled in PGS version 6.6. Please contact the Partek Licensing Team at licensing@partek.com to request this version or use the **Help > Check for Updates** command to check whether you have the latest released version. The screenshots shown within this tutorial may vary across platforms and across different versions of PGS.

Description of the Data Set

In this tutorial, you will analyze an RNA-Seq experiment using PGS's RNA-Seq workflow. The data used in this tutorial was generated from mRNA extracted from four diverse human tissues (skeletal muscle, brain, heart, and liver) from different donors and sequenced on the Illumina® Genome Analyzer™. The single-end mRNA-Seq reads were mapped to the human genome (hg19), allowing up to two mismatches, using Partek® Flow™ alignment and the default alignment options. The output files of Flow are BAM files which can be imported directly into Genomics Suite 6.6. BAM or SAM files from other alignment programs like ELAND (CASAVA), Bowtie, BWA, or TopHat are also supported. This same workflow will also work for aligned reads from any sequencing platform in the (aligned) BAM or SAM file formats.

Data and associated files for this tutorial can be downloaded by going to **Help > On-line Tutorials** from the PGS main menu.

Workflow

Open the RNA-Seq workflow within PGS by selecting it from the *Workflow* drop-down list in the upper right corner of the software. The entire workflow is shown in Figure 1.

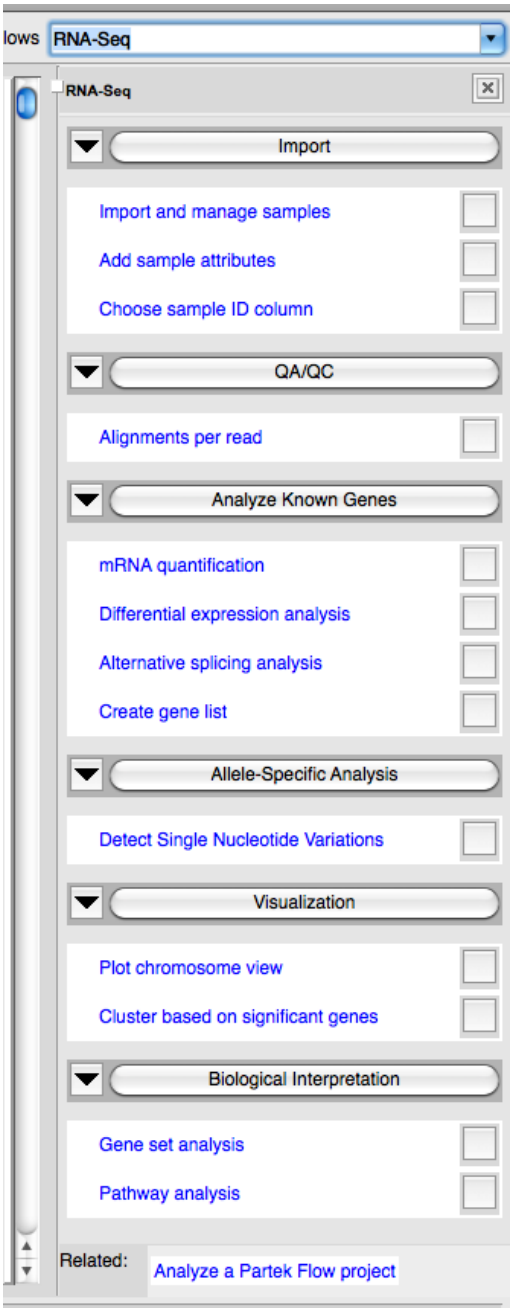


Figure 1: The RNA-Seq workflow

The *RNA-Seq* workflow will be used throughout this tutorial to analyze RNA-Seq data. These and other commands for analyzing RNA-Seq data are also available from the command toolbar.

Step 1- Importing the aligned reads

PGS can import large files (several gigabytes) of reads that are already aligned to a reference genome. The data used for this tutorial was already aligned by Partek® Flow™. The sequence importer can handle the two standard alignment formats, .BAM and .SAM. Conversion from ELAND .txt files to BAM files is available via the *Tools* menu. Also note that if a quantification project has been created in Partek Flow, this project can also be imported and analyzed; in this scenario, invoke the workflow from the very bottom of the standard *RNA-Seq* workflow (Figure 1: *Related: Analyze a Partek Flow project*).

- Under *Import* from the *RNA-Seq* workflow, select **Import and manage samples**. The *Sequence Import* window will open (Figure 2)

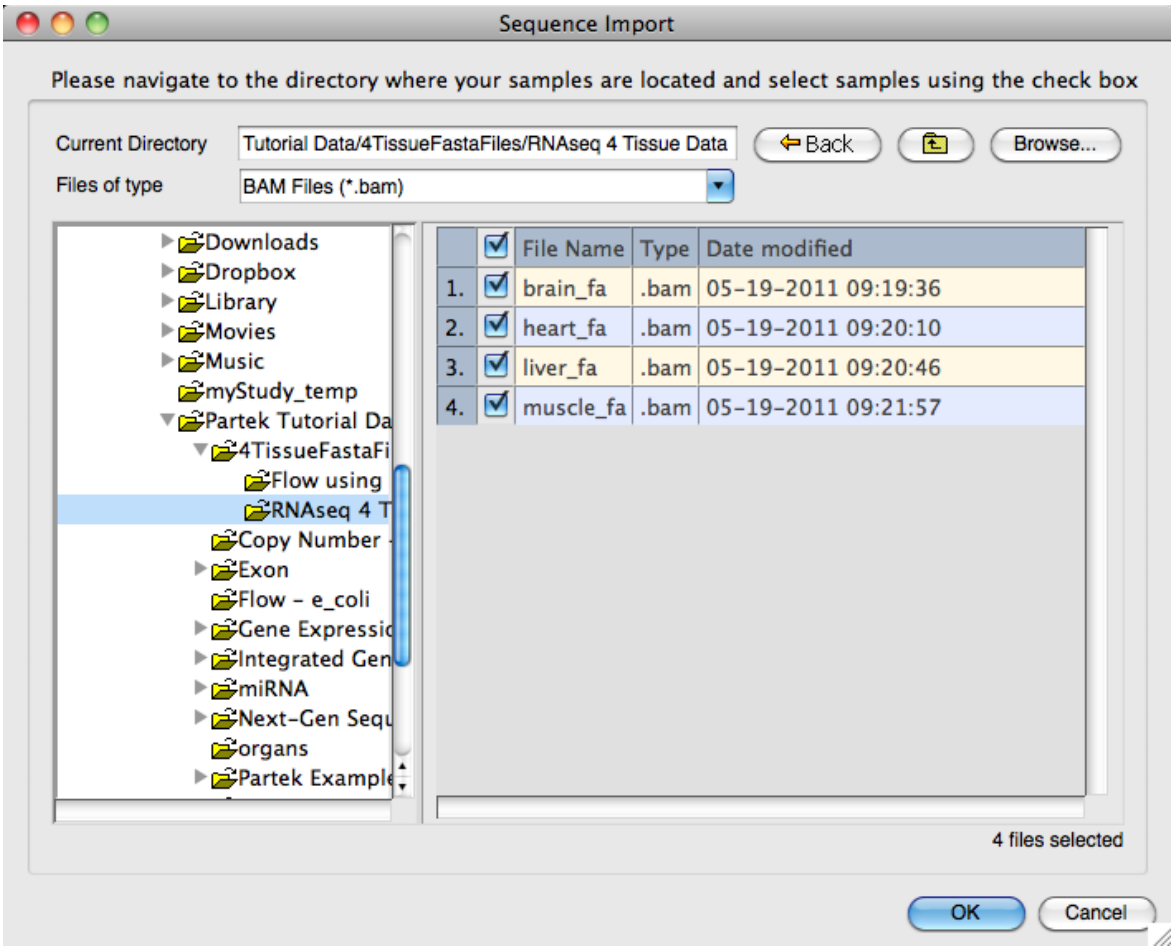


Figure 2: Viewing the Sequence Importer window. Choose the files you want to import

- *Files of type*: Select **BAM Files (*.bam)** from the drop-down list.
  Browse to the folder where you have stored the BAM files. Select the files to import by checking the box to the left of the data files. For this tutorial, select **brain_fa**, **heart_fa**, **liver_fa**, and **muscle_fa**
- Select **OK** and the next *Sequence Impor*t window will open as shown in Figure 3

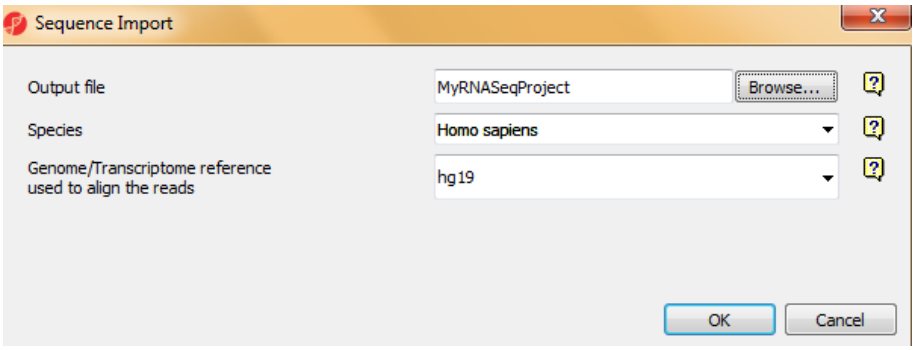


Figure 3: Viewing the Sequence Import wizard; specify Output file (and directory using Browse), Species, and Genome

Configure the dialog as follows:

- *Output file*: provide a name for the top-level spreadsheet. Use the *Browse* button to change the output directory
- *Species*: Select **Homo sapiens** from the drop-down menu since this data is from human subjects
- *Genome/Transcriptome reference used to align the reads*: Select the genome build against which your data was aligned. For this tutorial data, choose **hg 19** since the data was aligned to the reference genome hg19
- Select **OK**

This will open the *BAM Sample Manager* window (Figure 4). The *Bam Sample Manager* window shows the files to be imported. The *Manage sequence names* option allows you to check or modify the chromosome name mapping (not an issue for human samples but may cause problems with esoteric organisms if the chromosome names used by the aligner do not match the chromosome names in the genome annotations). It is also possible to add the samples to a separate experiment, rather than adding them to the existing experiment, by using the *Add new experiment* button. Samples may be removed from the experiment with *Remove selected samples;* samples must first be selected by clicking on the row in the list of samples.

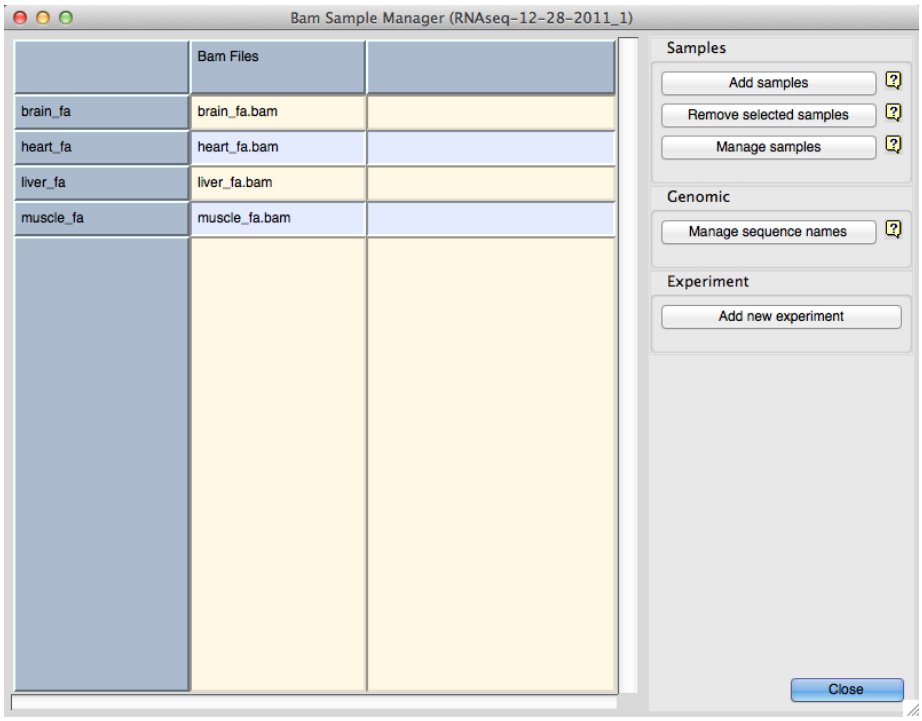


Figure 4: Bam Sample Manager window

- In this tutorial, the individual file names are short, but in some cases, the names may be much longer. Assigning shorter or more informative names will lead to clearer labels/legends later in the workflow, To change the names, select **Manage samples** to invoke the *Assign files to samples* dialog (Figure 5)
- The path to the file is shown, and the *Sample ID* is the filename by default. In this example, change the first sample to **Brain** (circled area)

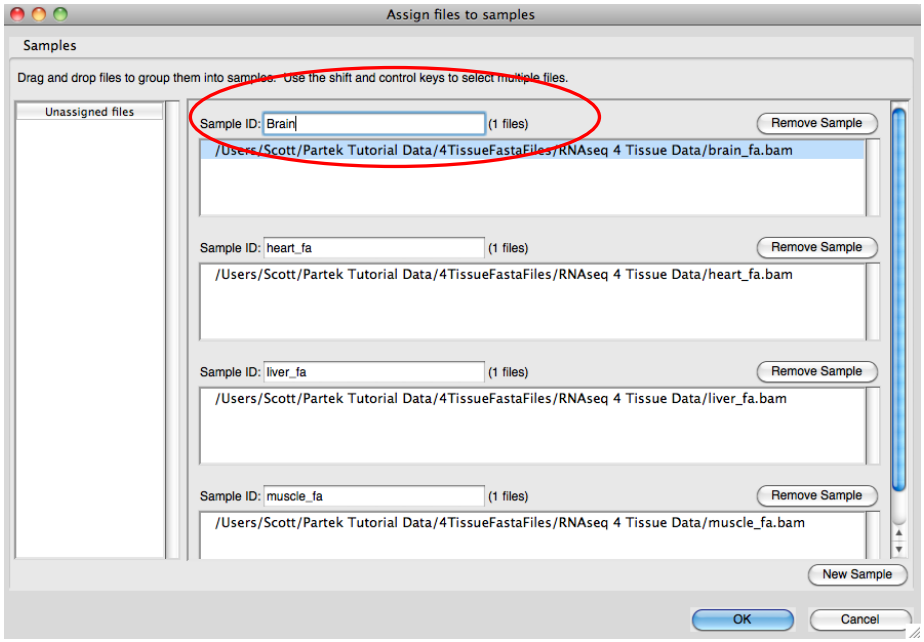


Figure 5: Using the Assign files to samples dialog to give informative sample names

- For this tutorial, the default names are suitable, so select **Cancel** to proceed. However, if you have data from one sample which is split into two or more BAM files, you can also use *Manage samples* to merge these reads into one sample. Additionally, you may use *Add samples* or *Remove selected samples* if needed
- Click **Close** to proceed. If the files are being imported for the first time, they will be sorted in order to quickly visualize and analyze the data

The imported data will appear in a spreadsheet (Figure 6). Each sample is listed in a row with the number of aligned reads displayed for each sample.

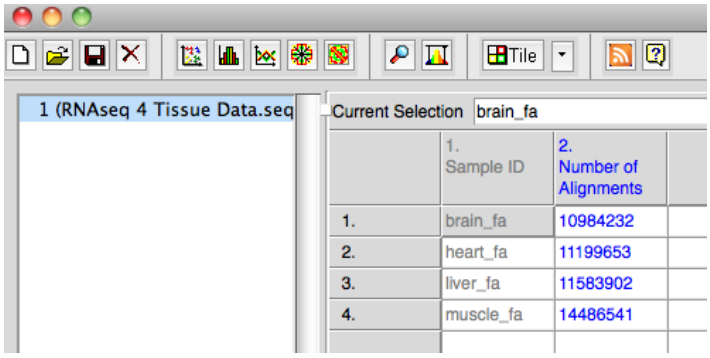


Figure 6: Viewing the imported data in a spreadsheet

Notice that the *Sample ID* names in column 1 are gray, denoting a text field, but we want to change these such that the sample names are a categorical factor that can be used in the downstream analysis. To change the properties of column 1, right-click on the column header to invoke the contextual menu and select **Properties**. In the resulting dialog, (Figure 7) set *Type* to **categorical**, *Attribute* to **factor**, and then select **OK**.

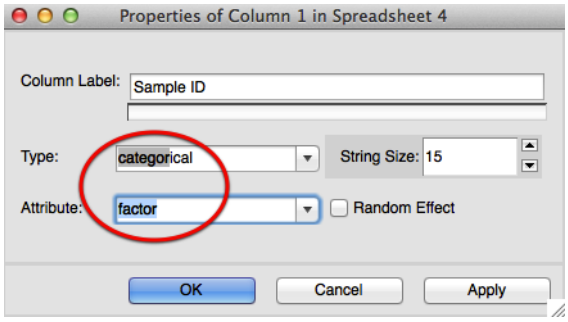


Figure 7: Setting column properties

The sample names in column 1 will now be black to denote a categorical variable. Next, we will add attributes for grouping the data, i.e., into replicates or sample groups. From the workflow, select **Add sample attribute,** choose the **Add a categorical attribute** option, and select **OK** (Figure 8)

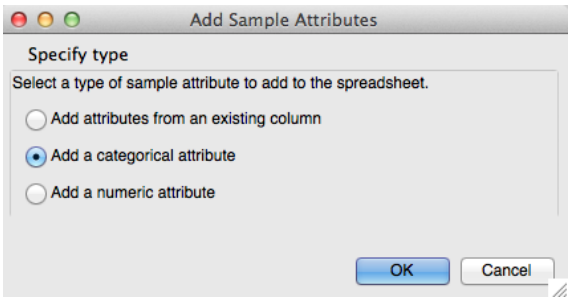

Add Sample Attributes
Specify type
Select a type of sample attribute to add to the spreadsheet.
Add attributes from an existing column
Add a categorical attribute
Add a numeric attribute
OK
Cancel

Figure 8: Adding sample attributes to describe or group samples by categories

In this tutorial, we have four samples from different tissues, but to illustrate the statistical analysis options later in the workflow, we will group the tissues into two groups: *muscle* (muscle and heart samples) and *NOT muscle* (liver and brain samples). These two groups will then be compared at a later step.

- In the dialog in Figure 9, change *Attribute name* to **Tissue**
- Rename *Group_1* to **Muscle** and *Group_2* to **NOT muscle**
- Drag the samples to the correct windows. The setup is shown in Figure 9. Additional groups can be added as required using the *New Group* button
- Select **OK** to proceed

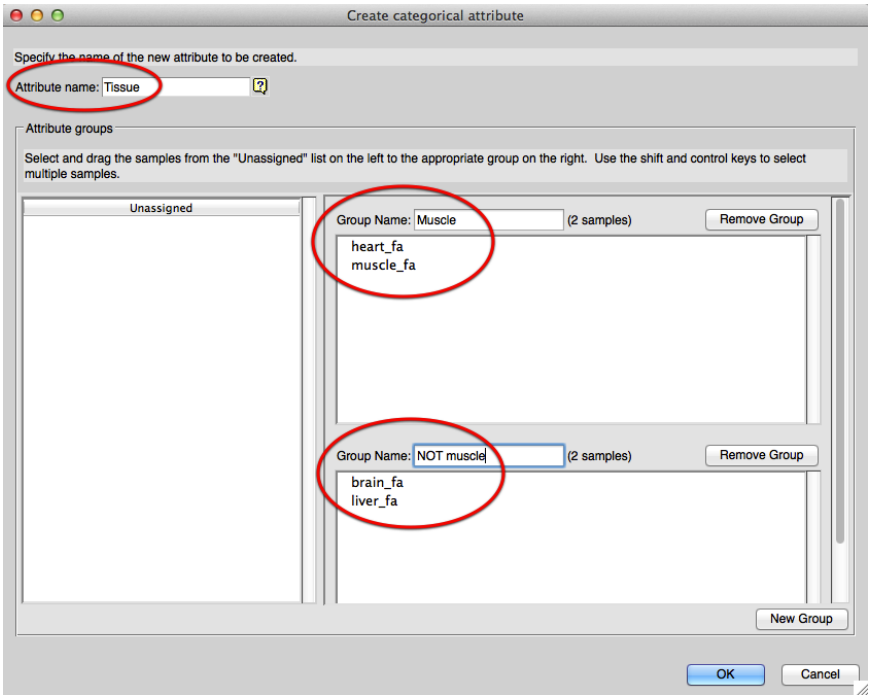


Figure 9: Create categorical attribute

- When asked if you wish to add another categorical attribute, select **No**, and if you wish to save the spreadsheet, select **Yes**. The attribute will now appear as a new column with the heading *Tissue* and the groups *Muscle* and *NOT muscle* (Figure 10)

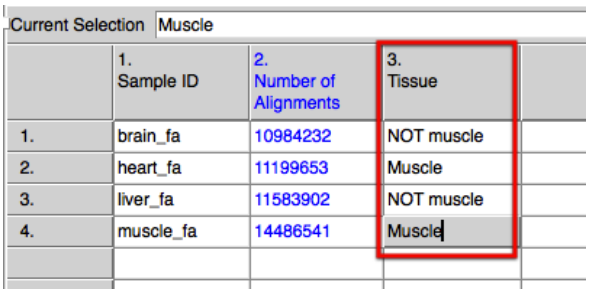


Figure 10: Tissue, a new categorical attribute, has been added

It is also important to ensure that the correct column is defined as the Sample ID. This is particularly important when integration of data from different experiments is desired.

- From the *Import* section of the workflow, select **Choose sample ID** column
- Use the drop-down menu to select the **Sample ID** column; the samples names will be shown (Figure 11)
- Select **OK**

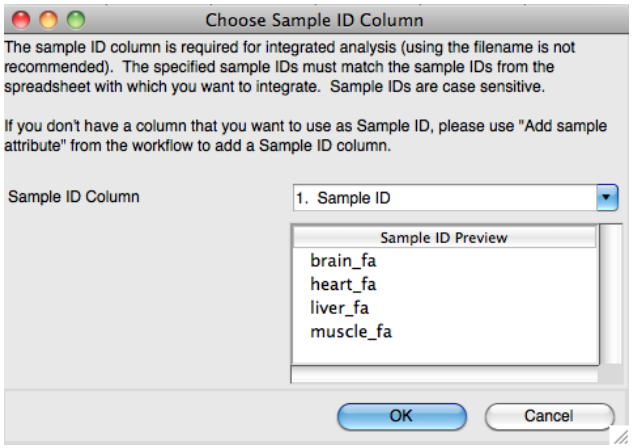

Choose Sample ID Column
The sample ID column is required for integrated analysis (using the filename is not recommended). The specified sample IDs must match the sample IDs from the spreadsheet with which you want to integrate. Sample IDs are case sensitive.
If you don't have a column that you want to use as Sample ID, please use "Add sample attribute" from the workflow to add a Sample ID column.
Sample ID Column
1. Sample ID
Sample ID Preview
brain_fa
heart_fa
liver_fa
muscle_fa
OK
Cancel

Figure 11: Specifying the correct Sample ID column

- For quality assessment, select **Alignments per read** from the *QA/QC* portion of the workflow. After analyzing the four samples, a new child spreadsheet named *Alignment_Counts* is created (Figure 12)

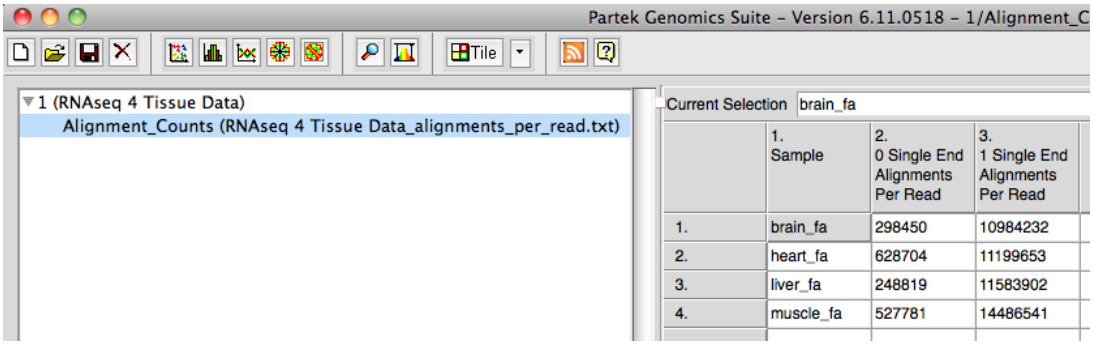

Partek Genomics Suite - Version 6.11.0518 - 1/Alignment_C
Tile
1 (RNAseq 4 Tissue Data)
Alignment_Counts (RNAseq 4 Tissue Data_alignments_per_read.txt)
Current Selection
brain_fa
1. Sample
2. 0 Single End Alignments Per Read
3. 1 Single End Alignments Per Read
1. brain_fa 298450 10984232
2. heart_fa 628704 11199653
3. liver_fa 248819 11583902
4. muscle_fa 527781 14486541

Figure 12: Alignment_Counts spreadsheet

In column 3, one can see that all imported aligned reads from Figure 6 align exactly once to the human genome. This is dependent on the options used during the alignment process. For other data and alignment options, one might observe more than one alignment per read. In column 2, there are single-end reads with zero alignments per read reported because BAM files also contain all the reads that were not aligned during the alignment process.

Step 2 – Visualization

Once imported, it is possible to visualize the mapped reads along with gene annotation information and cytobands.

- Select the parent spreadsheet (*RNA-Seq 4 Tissue Data* in this example)
- Select **Plot Chromosome View** in the *Visualization* section of the workflow. Unless you have previously downloaded an annotation file (during another experiment), you will be prompted to select an annotation source (Figure 13)

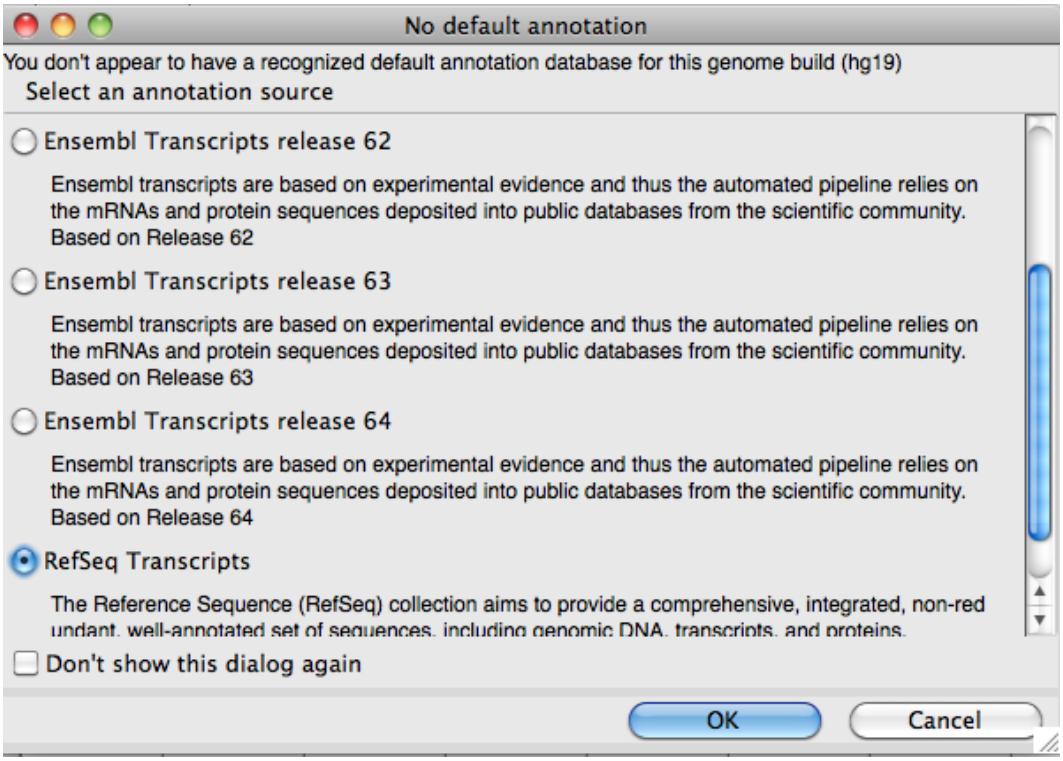

No default annotation
You don't appear to have a recognized default annotation database for this genome build (hg19)
Select an annotation source
Ensembl Transcripts release 62
Ensembl transcripts are based on experimental evidence and thus the automated pipeline relies on the mRNAs and protein sequences deposited into public databases from the scientific community. Based on Release 62
Ensembl Transcripts release 63
Ensembl transcripts are based on experimental evidence and thus the automated pipeline relies on the mRNAs and protein sequences deposited into public databases from the scientific community. Based on Release 63
Ensembl Transcripts release 64
Ensembl transcripts are based on experimental evidence and thus the automated pipeline relies on the mRNAs and protein sequences deposited into public databases from the scientific community. Based on Release 64
RefSeq Transcripts
The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-red
undant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.
Don't show this dialog again
OK
Cancel

Figure 13: Selecting an annotation source

- For this tutorial, select **RefSeq Transcripts**. PGS will download the relevant file (subject to any firewall/proxy settings) and save it to your default library location. You may also notice a dialog flash as the cytoband file is automatically downloaded. The *Genome Viewer* window will open with chromosome 1 displayed (Figure 14). You may choose other chromosomes from the drop-down menu (circled) to change which chromosome is displayed

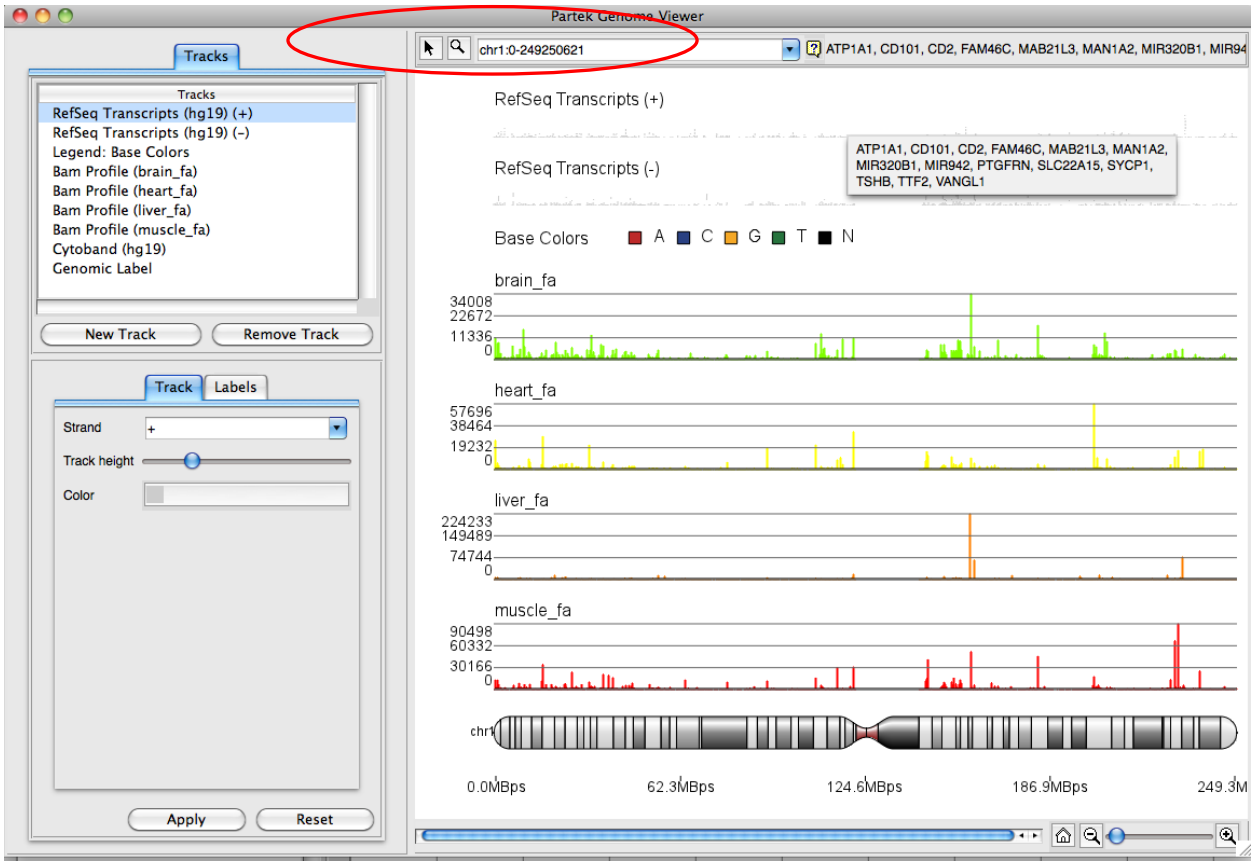


Figure 14: Visualizing reads on a chromosome level in the Genome View

The panel on the left shows the ten tracks in the viewer. The *New Track* button adds a new track into the viewer (Figure 15), and the *Remove Track* button removes the selected track from the viewer.

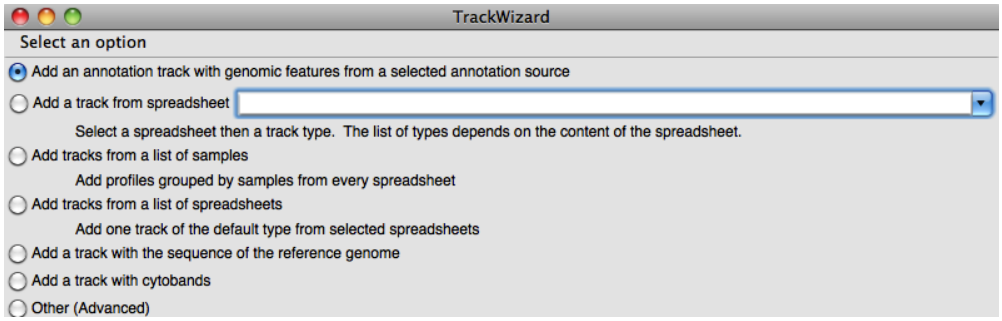

TrackWizard
Select an option
Add an annotation track with genomic features from a selected annotation source
Add a track from spreadsheet
Select a spreadsheet then a track type. The list of types depends on the content of the spreadsheet.
Add tracks from a list of samples
Add profiles grouped by samples from every spreadsheet
Add tracks from a list of spreadsheets
Add one track of the default type from selected spreadsheets
Add a track with the sequence of the reference genome
Add a track with cytobands
Other (Advanced)

Figure 15: Adding a new track to the Genome Viewer

The following is a detailed description of each track in the default view.

RefSeq Transcripts (+)

The *RefSeq Transcripts (+)* track shows all genes encoded on the forward strand of chromosome 1. This experiment uses RefSeqGene, which defines genomic sequences of well-characterized genes, to be used as the reference annotation track. Mouse-over a

particular region in this track, and all genes within this region are shown in the information bar (visible in the top right of Figure 14). Zoom into this track to see individual genes, including alterative isoforms. Zooming in on one track automatically zooms all other visible tracks, thus you can now see the reads that mapped to this particular gene across all samples.

RefSeq Transcripts (-)

The *RefSeq Transcripts (-)* track shows all the genes encoded on the reverse strand of chromosome 1.

Legend: Base Colors

This track indicates the color coding for individual bases. Although included in the default view, the individual bases are only visible once zoomed into a region of interest.

Bam Profile (muscle_fa, brain_fa, heart_fa, and liver_fa)

These tracks show all the reads that mapped to chromosome one from the four tissue samples. The y-axis numbers on the left side of the tracks indicate the raw read counts. The aligned reads are shown in the *Genome Viewer* in each track with a different color for each *Bam Profile* track.

Genome Sequence, Cytoband and Genomic Label

The *Genome Sequence*, *Cytoband* and *Genomic Label* tracks are shown at the bottom of the panel. *Genome Sequence* will display the bases (different colors and labels) of the reference genome specified when zoomed in sufficiently. The other two labels are helpful for navigating about the chromosome.

Using the Genome Viewer

In the *Genome Viewer*, select for selection mode and for navigation mode. In navigation mode, to zoom in on a region of interest, left-click and draw a box to zoom in on any region for a detailed view of the reads mapped to this region. Alternatively, in the bottom left-hand corner of the browser, drag the slider left-to-right or click the plus (+) button to zoom in. Drag right-to-left or click the minus (-) button to zoom back out.

It is also possible to zoom directly to a specific gene of interest using *Genome View*. For example, type **MYC** into the search box at the top of the window, and the viewer will show just the MYC transcript in the *RefSeq* track and the aligned reads (Figure 16). The individual base colors are now visible. To return to the whole chromosome view, simply select .

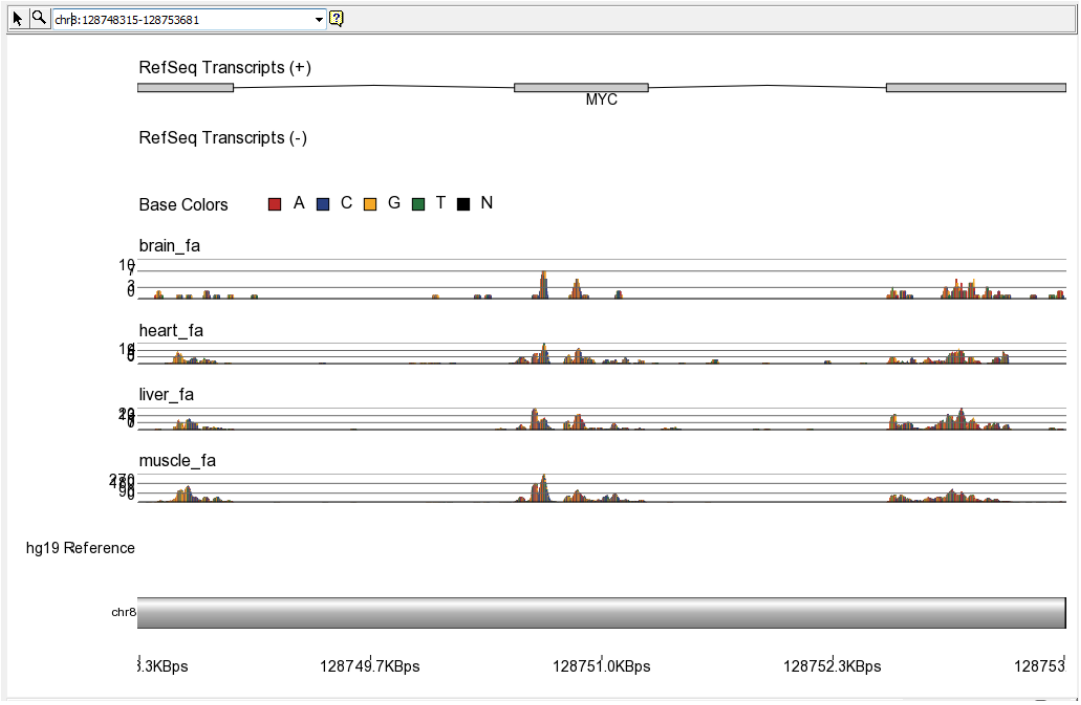


Figure 16: Zooming to look at a gene of interest (MYC)

Step 3 - Analyze Known Genes

The next step in the workflow is to detect differentially expressed genes by performing **mRNA quantification**. This function creates data tables at the transcript and gene levels and identifies those transcripts that are differentially expressed or spliced across all samples. The raw and normalized reads are also reported for each sample. The normalization method used by PGS is Reads Per Kilobase of exon model per Million mapped reads (RPKM) (Mortazavi et al.).

- Select **mRNA quantification** from the *Analyze Known Genes* section of workflow

The dialog box shown in Figure 17 will appear. Your choices for these options depend on the aims of your experiment. In the *Configure the test* section (Figure 17), you are asked if the assay can discriminate between the sense and antisense transcripts. Your answer depends on the method used for sample preparation as some methods preserve the strand information of the original transcript and some do not. A directional mRNA-Seq sample prep protocol only synthesizes the first strand of cDNA whereas other methods reverse transcribe the mRNA into double-stranded cDNA. In the latter case, the sequencer reads sequences from both the sense and antisense strands but does not discriminate between them.

- Select **No** for *Can assay discriminate between sense and antisense transcripts?* because the library preparation method did not preserve the strand information

The dialog also asks if you would like the intronic reads to be compatible with the gene in the gene-level result spreadsheets. By selecting *Yes*, reads that might correspond to new or

extended exons will be considered part of the gene in the gene-level spreadsheets, and the RPKM calculation will include the intron length in the transcript length.

- For *In the gene-level result report intronic reads as compatible with the gene?*, select **No**

The next option *Report results with no reads from any sample?* determines whether the results spreadsheets include transcripts or genes that have no reads from any sample. Selecting *Yes* will include all the genes/transcripts in the transcriptome even if there are no reads for that transcript/gene from any sample.

- As this tutorial is not concerned with genes that are not expressed in all samples, select **No** for *Report results with no reads from any sample?*

Lastly, *Report unexplained regions with more than reads* determines if reads that are mapped to the genome but not to any transcript are reported on a separate spreadsheet (*unexplained_regions)*. If checked, you must also specify the minimum number of reads that must be present before the region is reported.

- Make sure **Report unexplained regions with more than reads** is checked and specify **5** as the number of reads

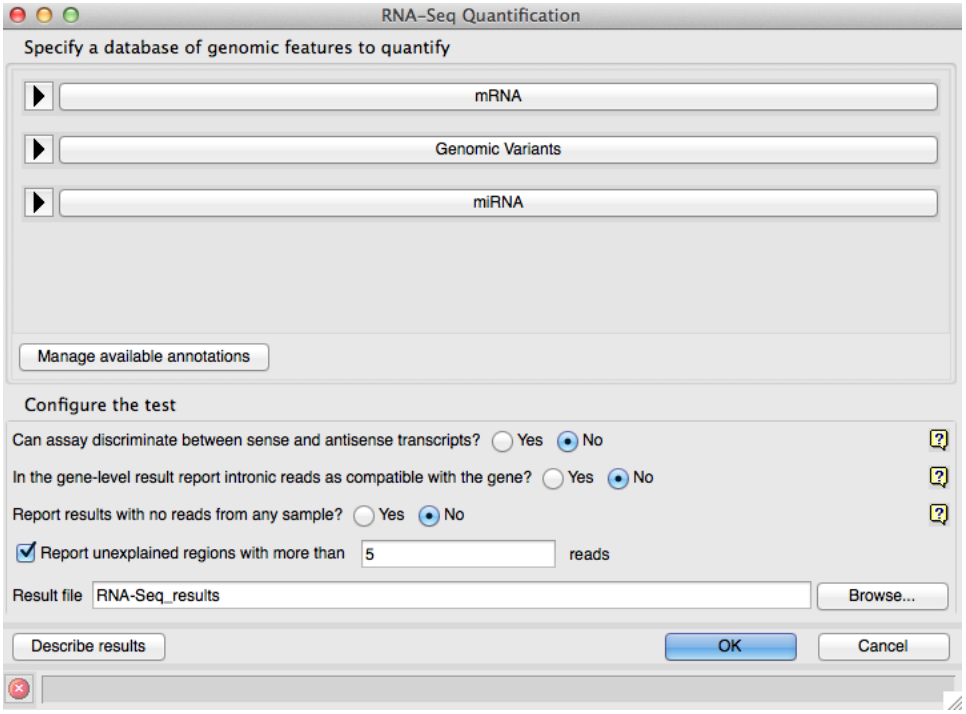

RNA-Seq Quantification
Specify a database of genomic features to quantify
mRNA
Genomic Variants
miRNA
Manage available annotations
Configure the test
Can assay discriminate between sense and antisense transcripts? Yes No
In the gene-level result report intronic reads as compatible with the gene? Yes No
Report results with no reads from any sample? Yes No
Report unexplained regions with more than 5 reads
Result file RNA-Seq_results
Browse...
Describe results
OK
Cancel

Figure 17: RNA-Seq Quantification options

- In this tutorial, you will use an mRNA database, RefSeq. Select the black arrowhead (circled in Figure 18), select **RefSeq Transcripts** and **OK** to continue

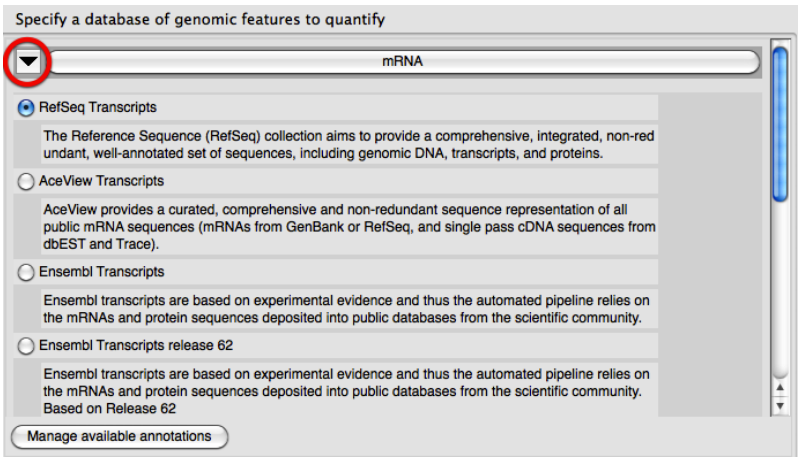

Specify a database of genomic features to quantify
mRNA
RefSeq Transcripts
The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.
AceView Transcripts
AceView provides a curated, comprehensive and non-redundant sequence representation of all public mRNA sequences (mRNAs from GenBank or RefSeq, and single pass cDNA sequences from dbEST and Trace).
Ensembl Transcripts
Ensembl transcripts are based on experimental evidence and thus the automated pipeline relies on the mRNAs and protein sequences deposited into public databases from the scientific community.
Ensembl Transcripts release 62
Ensembl transcripts are based on experimental evidence and thus the automated pipeline relies on the mRNAs and protein sequences deposited into public databases from the scientific community. Based on Release 62
Manage available annotations

Figure 18: Specifying a transcriptome database

Reads will now be assigned to individual transcripts of a gene based on the Expectation/Maximization (E/M) algorithm (Xing, et al.). In PGS, the E/M algorithm is modified to accept paired-end reads, junction aligned reads, and multiple aligned reads if these are present in your data. For a detailed description of the E/M algorithm, refer to the RNA-Seq white paper (*Help > On-line Tutorials > White Papers).*

- Select **OK** to perform the RNA-Seq analysis, which will generate a mapping summary spreadsheet and up to six other spreadsheets containing the analyzed results (Figure 19)

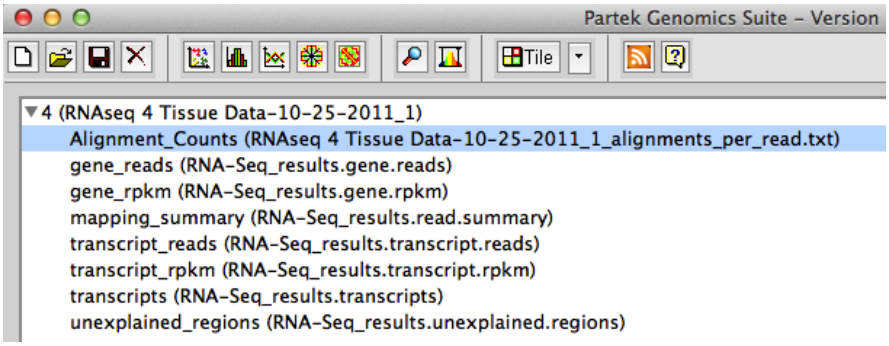


Figure 19: Viewing the list of results spreadsheets after detecting differential expression and alternative splicing

The same dialog box that appears if you select *Describe results* will pop-up if you have not previously turned this off. This dialog box explains the spreadsheets generated by *mRNA quantification.*

The *mapping_summary* spreadsheet

The *mapping_summary* spreadsheet (Figure 20) is a summary of reads that have been mapped to the intronic, xxonic or intergenic regions. This is a QA/QC step to give you a brief idea of how the reads are distributed across the transcriptome. You should expect the results of replicates to resemble each other with respect to the distribution of reads.

Current Selection: brain_fa

	1. Sample ID	2. Number of Alignments	3. Tissue	4. Total Number of Reads	5. Percentage of reads which fully overlap exon	6. Percentage of reads which partially overlap exon	7. Percentage of reads within an intron	8. Percentage of reads between genes
1.	brain_fa	10984232	NOT muscle	10984232	61.9032	1.59361	5.62868	30.8745
2.	heart_fa	11199653	Muscle	11199653	48.8748	1.77955	6.79877	42.5468
3.	liver_fa	11583902	NOT muscle	11583902	76.8934	2.026	4.27321	16.8073
4.	muscle_fa	14486541	Muscle	14486541	68.3527	2.34444	5.46389	23.839

Figure 20: Viewing the RNA-Seq_results.read.summary spreadsheet

The *gene_reads* and *gene_rpkm* spreadsheets

PGS presents the gene-level data both before (reads) and after (RPKM) normalization. Samples are listed one per row with the normalized counts of the reads mapped to the genes in columns. As shown in Figure 21, you can see that there are 18,971 columns representing 18,968 genes. The normalization is by RPKM (Mortazavi, et al.).

The *gene_rpkm* spreadsheet is particularly useful when you have biological replicates in your sample groups. You may go to *View* > *Scatter Plot* from the PGS toolbar to create a PCA plot and examine how your samples group together. For a detailed introduction of PCA, refer to Chapter 7 of the Partek User's Manual. With replicates, you would also be able to perform *Differential expression analysis* using ANOVA with the *gene_rpkm* spreadsheet.

Current Selection: 0

	1. Sample ID	2. Number of Alignments	3. Tissue	4. A1BG	5. A1BG-AS1	6. A1CF	7. A2LD1	8. A2M	9. A2ML1	10. A4GALT
1.	brain_fa	10984232	Not muscle	0.849542	0.922038	0	0.764269	23.8571	0.508895	1.39258
2.	heart_fa	11199653	Muscle	0.606717	0.249575	0	0.227591	106.411	0	20.1881
3.	liver_fa	11583902	Not muscle	438.696	14.4031	10.9251	2.33816	551.183	0	0.825303
4.	muscle_fa	14486541	Muscle	0.273617	0	0.014783	0.297884	24.3486	0	15.5086

Rows: 4 Cols: 18971

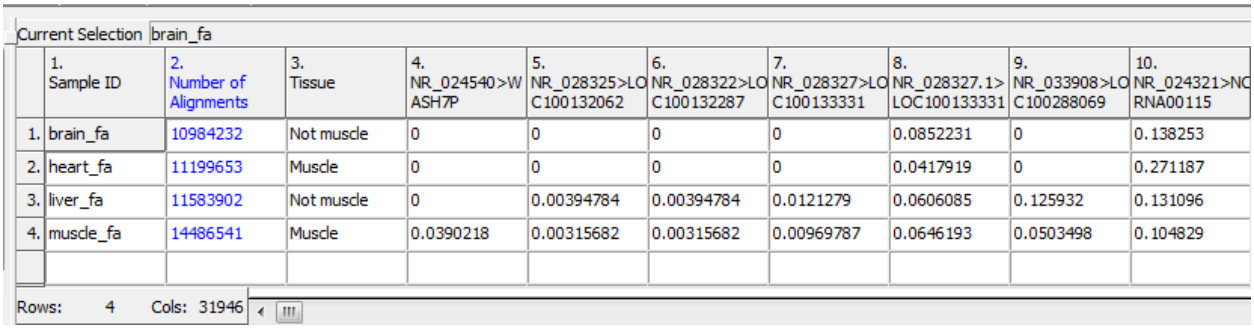
Figure 21: Viewing the RNA-Seq_result.gene.rpkm spreadsheet

The *transcript_reads* and *transcript_rpkm* spreadsheets

As above, PGS presents the transcript-level data both before and after normalization. The normalized count of sequencing reads are mapped to each transcript, listed as one sample per row with transcript IDs in columns. As shown in Figure 22, you can see that there are 31,946 columns representing 31,943 annotated transcripts. The normalization is by RPKM (Mortazavi, et al.).

If you have biological replicates in your sample groups and want to do differential expression on the transcript level, this is the spreadsheet to you would use. Similarly, if you have biological replicates in your sample groups and want to perform *Alternative splicing analysis*, this is the spreadsheet to use as input. The same analysis (PCA and ANOVA) can

be done on this spreadsheet as described above in the *RNA-Seq_result.gene.rpkm* spreadsheet.



Current Selection: brain_fa

	1. Sample ID	2. Number of Alignments	3. Tissue	4. NR_024540>WASH7P	5. NR_028325>LOC100132062	6. NR_028322>LOC100132287	7. NR_028327>LOC100133331	8. NR_028327.1>LOC100133331	9. NR_033908>LOC100288069	10. NR_024321>NCRNA00115
1.	brain_fa	10984232	Not muscle	0	0	0	0	0.0852231	0	0.138253
2.	heart_fa	11199653	Muscle	0	0	0	0	0.0417919	0	0.271187
3.	liver_fa	11583902	Not muscle	0	0.00394784	0.00394784	0.0121279	0.0606085	0.125932	0.131096
4.	muscle_fa	14486541	Muscle	0.0390218	0.00315682	0.00315682	0.00969787	0.0646193	0.0503498	0.104829

Rows: 4 Cols: 31946

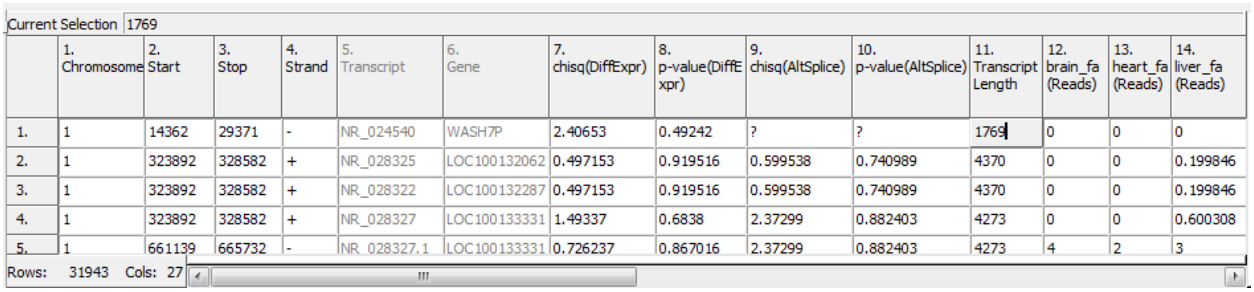
Figure 22: Viewing the RNA-Seq_result.transcript.rpkm spreadsheet

The *transcripts* spreadsheet

The *transcripts* spreadsheet details the analysis results of RNA-Seq if there no replicates (Figure 23 and Figure *24*). Each row lists a separate transcript. A description of each column can be found in Table 1.

Table 1: Column Descriptions in the transcripts spreadsheet

Column Label	Description
Chromosome, Start, Stop	Genomic location of the transcript
Strand, Transcript, Gene	Information about the transcript and gene symbol; + and - indicate the transcript is transcribed from the forward or reverse strand, respectively; transcript name is determined by the transcriptome database selected (NCBI mRNA ID in this case)
chisq and p-value (DiffExpr)	In the absence of replicates, differences in expression between samples are detected using a log-likelihood test at the transcript level. Sorting by ascending p-value shows the most significant differentially-expressed transcript at the top of the spreadsheet
chisq and p-value (AltSplice)	In the absence of replicates, splice-variants are detected using a Pearson chi-square test at the gene level. Sorted by ascending p-value shows genes with the greatest probability of expressing different RNA isoforms between samples at the top of the spreadsheet
Transcript Length	Length of transcript in base pairs (bp)
Read counts (raw) of each sample	Number of reads directly from the aligned data assigned to the transcript for each sample
Read counts (RPKM) of each sample	Raw read counts normalized by RPKM (Mortazavi, et al.) assigned to the transcript for each sample
Junction RPKM	If the aligner that was used to align this data supported junction reads, these columns (one per sample) will show the normalized RPKM reads assigned to junction reads for this transcript
Incompatible RPKM	Paired reads only; normalized unique paired reads (RPKM) that intersect with the transcript but are considered incompatible because the mate is not found in the same transcript

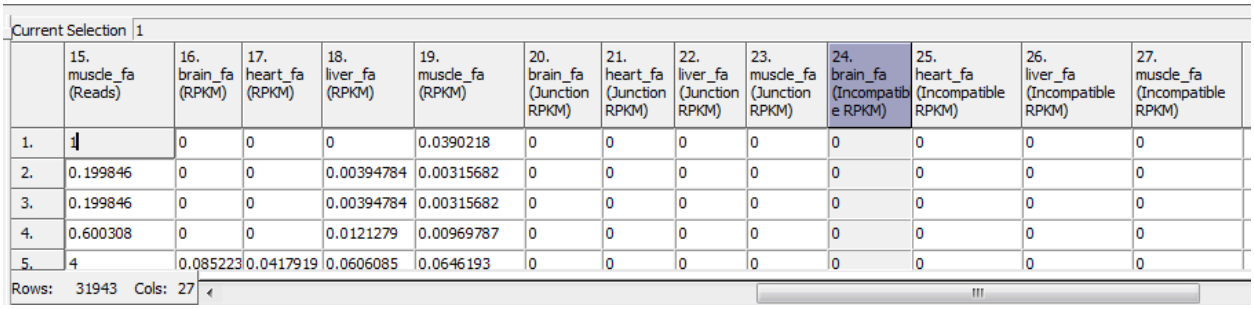


Current Selection 1769

	1. Chromosome	2. Start	3. Stop	4. Strand	5. Transcript	6. Gene	7. chisq(DiffExpr)	8. p-value(DiffExpr)	9. chisq(AltSplice)	10. p-value(AltSplice)	11. Transcript Length	12. brain_fa (Reads)	13. heart_fa (Reads)	14. liver_fa (Reads)
1.	1	14362	29371	-	NR_024540	WASH7P	2.40653	0.49242	?	?	1769	0	0	0
2.	1	323892	328582	+	NR_028325	LOC100132062	0.497153	0.919516	0.599538	0.740989	4370	0	0	0.199846
3.	1	323892	328582	+	NR_028322	LOC100132287	0.497153	0.919516	0.599538	0.740989	4370	0	0	0.199846
4.	1	323892	328582	+	NR_028327	LOC100133331	1.49337	0.6838	2.37299	0.882403	4273	0	0	0.600308
5.	1	661139	665732	-	NR_028327.1	LOC100133331	0.726237	0.867016	2.37299	0.882403	4273	4	2	3

Rows: 31943 Cols: 27

Figure 23: Viewing the first part of the RNA-Seq_result.transcripts spreadsheet



Current Selection 1

	15. muscle_fa (Reads)	16. brain_fa (RPKM)	17. heart_fa (RPKM)	18. liver_fa (RPKM)	19. muscle_fa (RPKM)	20. brain_fa (Junction RPKM)	21. heart_fa (Junction RPKM)	22. liver_fa (Junction RPKM)	23. muscle_fa (Junction RPKM)	24. brain_fa (Incompatible RPKM)	25. heart_fa (Incompatible RPKM)	26. liver_fa (Incompatible RPKM)	27. muscle_fa (Incompatible RPKM)
1.	1	0	0	0	0.0390218	0	0	0	0	0	0	0	0
2.	0.199846	0	0	0.00394784	0.00315682	0	0	0	0	0	0	0	0
3.	0.199846	0	0	0.00394784	0.00315682	0	0	0	0	0	0	0	0
4.	0.600308	0	0	0.0121279	0.00969787	0	0	0	0	0	0	0	0
5.	4	0.085223	0.0417919	0.0606085	0.0646193	0	0	0	0	0	0	0	0

Rows: 31943 Cols: 27

Figure 24: Viewing the second part of the RNA-Seq_result.transcripts spreadsheet

It is possible to derive basic information about differential and alternative splicing between your samples if you don't have replicates from the *RNA-Seq_result.transcripts* spreadsheet using a simple chi-squared or log-likelihood tests since each sample is represented only once and the null hypothesis is that the transcripts are evenly distributed across all samples. However, the power of PGS resides in the implementation of a mixed-model ANOVA that can handle unbalanced and incomplete datasets, nested designs, numerical and categorical variables, any number of factors, and flexible linear contrasts when you do have biological replicates.

Detecting differential expression using the Partek® ANOVA

During import, you created a categorical attribute called *Tissue* and assigned the 4 samples to either the *Muscle* or *NOT muscle* groups. This step was to create replicates within a group, albeit this grouping is somewhat artificial and is used in this dataset simply to illustrate the ANOVA. Replicates are a prerequisite for analysis of differential expression using the ANOVA test.

- Select **Differential expression analysis** from the *Analyze Known Genes* section of the workflow. You have the choice of analyzing at either the *Gene*- or *Transcript-level*. Select **Gene-level analysis** (Figure 25)
- Make sure the **gene_rpkm** spreadsheet is selected as the *Spreadsheet*
- Select **OK** to open the ANOVA dialog

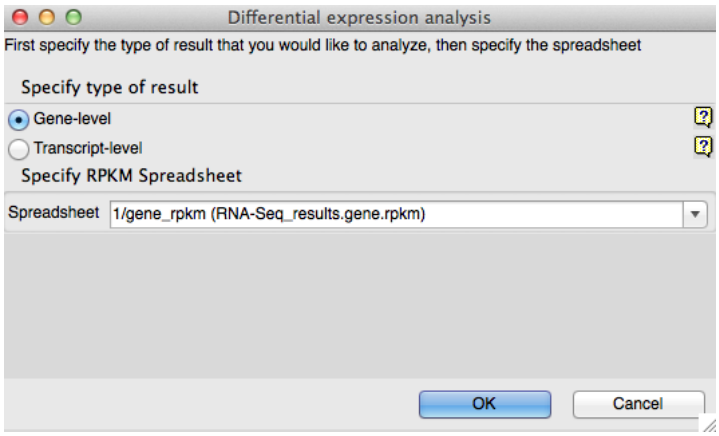


Figure 25: Differential expression analysis

- Categorical (experimental) factors are listed in the *Experimental Factor(s)* panel on the left. Move **Tissue** to the *ANOVA Factor(s)* pane on the right as shown in Figure 26

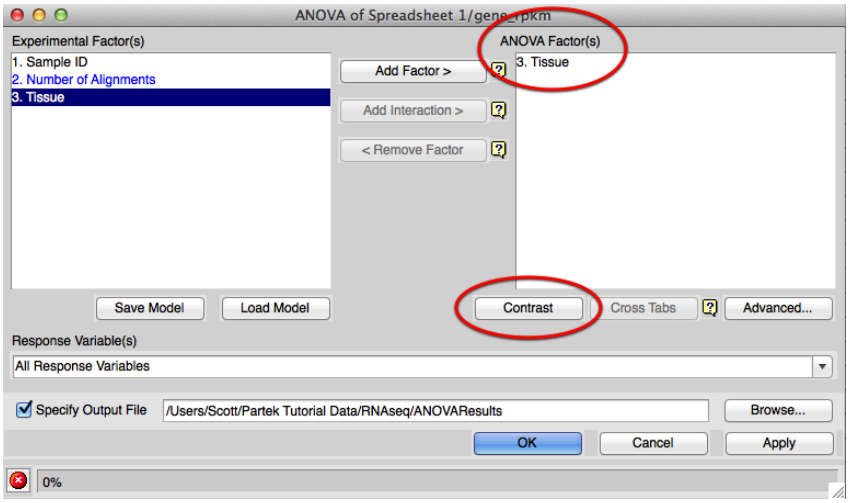


Figure 26: The ANOVA dialog

If the ANOVA were now performed (without contrasts), a p-value for differential expression would be calculated but would only indicate if there are differences within the factor *Tissue*; it will not inform you which groups are different or give any information on the magnitude of the change between groups (fold-change or ratio). To get this more specific information, you need to define linear contrasts.

- Select the **Contrast** button (circled in Figure 26)
- For *Select Factor/Interaction*, **Tissue** will be the only factor available as it was the only factor included in the ANOVA model in the previous step; if multiple factors were included, they could be selected here (top red circle)
- To define a contrast between the two candidate levels (**Muscle** and **NOT muscle**), select and move them as shown in Figure 27. The bottom group is the reference or control group. By defining this contrast, you will produce a specific p-value and a

fold change with the reference group as the denominator in the linear fold-change calculation

- Select **Add Contrast** (circled)
- Select **OK** to return to the *ANOVA* dialog and select **OK** again to perform the ANOVA

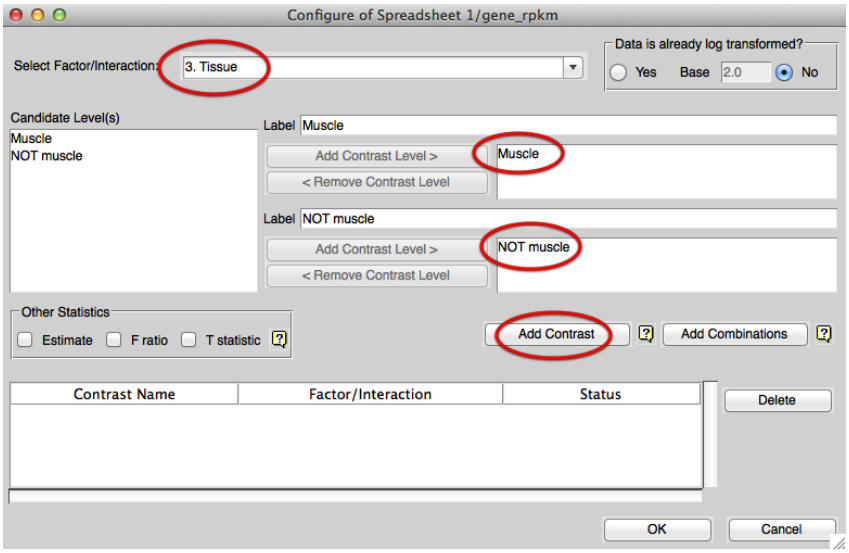


Figure 27: Define linear contrasts

Once the ANOVA has been performed on each gene in the dataset, an ANOVA child spreadsheet *ANOVA-1way (ANOVAResults)* will appear under the *gene_rpkm* sheet (Figure 28).

Current Selection 8876

	1. Column #	2. Column ID	3. Gene Symbol	4. p-value(Tissue)	5. p-value(Muscle vs. NOT muscle)	6. MeanRatio(Muscle vs. NOT muscle)	7. FoldChange(Muscle vs. NOT muscle)	8. FoldChange(Muscle vs. NOT muscle) (Description)	9. F(Tissue)	10. SS(Tissue)	11. SS(Error)	12. F(Error)
1.	6773	GPR98	GPR98	7.3945e-06	7.3945e-06	?	?	No change between Muscle and NOT	135234	0.301446	4.45814e-06	1
2.	17916	VWA1	VWA1	8.25054e-06	8.25054e-06	0.136816	-7.30908	Muscle down vs NOT muscle	121203	362.263	0.0059778	1
3.	14378	SCARA5	SCARA5	7.10032e-05	7.10032e-05	26.873	26.873	Muscle up vs NOT muscle	14082.4	73.2804	0.0104074	1
4.	16178	SYTL5	SYTL5	0.000145978	0.000145978	?	?	No change between Muscle and NOT	6848.84	0.418199	0.000122122	1
5.	16823	TMEM33	TMEM33	0.00014654	0.00014654	0.243243	-4.11112	Muscle down vs NOT muscle	6822.59	11.9366	0.00349913	1
6.	9181	LOC339751	LOC339751	0.000174524	0.000174524	?	?	No change between Muscle and NOT	5728.38	0.326635	0.000114041	1
7.	17804	VASN	VASN	0.000196221	0.000196221	0.368362	-2.71472	Muscle down vs NOT muscle	5094.81	19.7077	0.00773639	1
8.	679	ANKRD56	ANKRD56	0.000197364	0.000197364	0.00389606	-256.669	Muscle down vs NOT muscle	5065.28	9.12651	0.00360356	1
9.	8485	KRT74	KRT74	0.000197462	0.000197462	?	?	No change between Muscle and NOT	5062.78	13.2353	0.00522849	1
10.	11657	OGDHL	OGDHL	0.000212638	0.000212638	0.0143701	-69.5889	Muscle down vs NOT muscle	4701.33	234.535	0.099774	1

Figure 28: The ANOVA results sheet showing p-value, Mean Ratio, and Fold Change for each gene. The ? indicates that values could not be calculated (one group has no reads)

The format of the ANOVA spreadsheet is similar for all workflows. The description of each column may be found in Table 2.

Table 2: Interpretation of ANOVA results

Column Label	Description
Column #	The column number in the *gene_rpkm* spreadsheet
Column ID & Gene Symbol	Gene Symbol based on the annotation file used
p-value (Tissue)	The overall p-value for the *Tissue* factor included in the ANOVA model
p-value (Muscle vs. NOT muscle)	The specific p-value for the linear contrast of candidate levels within the factor (in this example there are only two levels)
Mean Ratio	The ratio of the mean values of RPKMs of the two groups selected in the contrast where Group_2 is the denominator (reference); numbers less than one imply down-regulation
FoldChange and FoldChange Description	Linear fold change (with 0 indicating NO change, negative numbers indicating down-regulation, and positive numbers indicating up-regulation) for each contrast defined and a text description
F(Tissue)	The F-statistic (essentially the ratio of signal-to-noise; high F-value = low p-value) for each factor
SS(Tissue)	The sum of squares for each factor (rough estimate of variability within the groups)
SS(Error) and F(Error)	Error is the variability in the data not explained by the factors included in the ANOVA (noise). The F-value is always set to 1 (ratio of noise-to-noise)

Note that in this tutorial, the overall p-value for the factor (column 4) is the same as the p-value for the linear contrast (column 5) as there are only two levels within factor. If we had more than two groups, the overall p-value and the linear contrast p-values would most likely differ. You can also see the ? symbol in the ratio/fold-change columns (6 and 7) for several genes that also have a low p-value, resulting from zero reads in one of the groups, thus ratios and fold-changes cannot be calculated.

For more detailed examples of setting up the ANOVA including multiple factors and linear contrasts, please refer to the gene expression tutorials (Down's Syndrome and Breast Cancer) available from *Help > On-line Tutorials*.

The *unexplained_regions* spreadsheet

The contents of this sheet will be explained in more detail in **Step** 8.

Step 4 – Use the Genome Browser

- Select the ***transcripts*** spreadsheet and then select **Plot Chromosome View** under the *Visualization* tab to view the analyzed results in Partek (Figure 29)

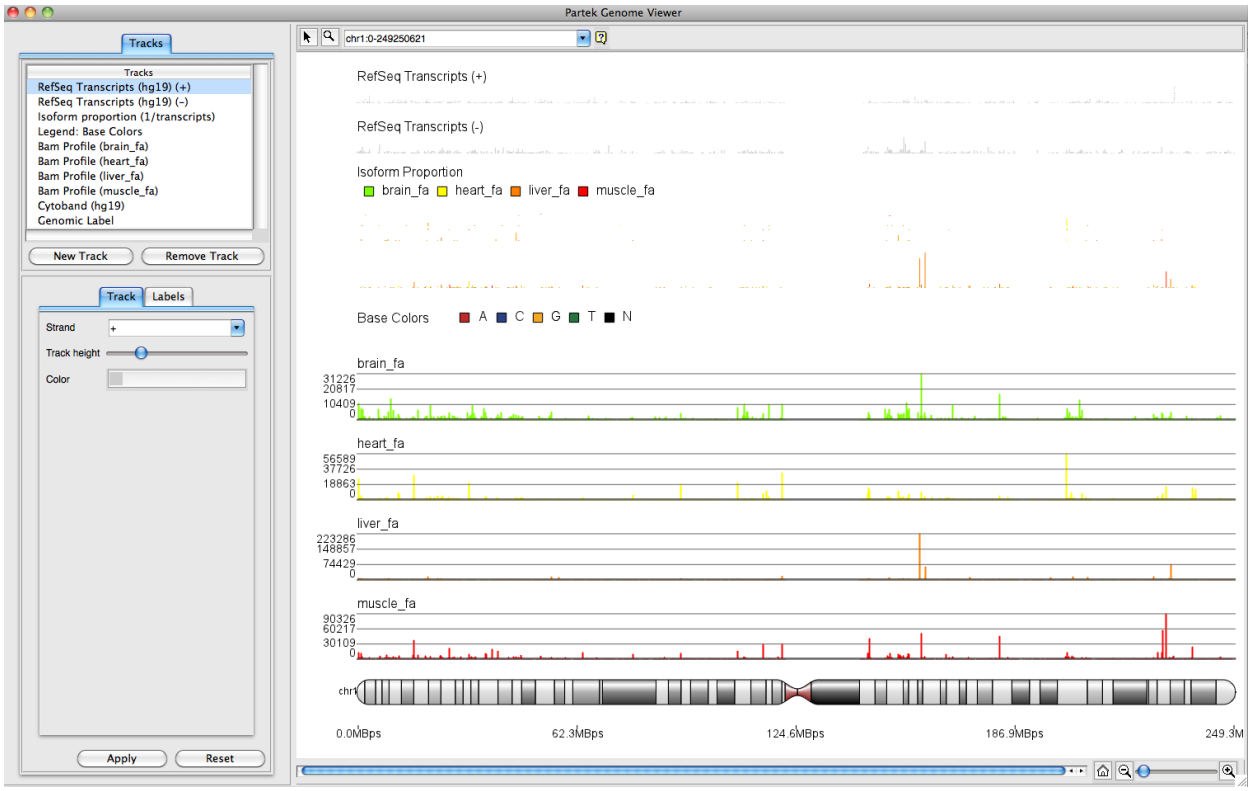


Figure 29: Viewing the transcript results in the Genome Viewer

The viewer is almost identical to that seen after importing the aligned reads; the difference is the inclusion of the Isoform Proportions track (highlighted in the *Tracks* list in Figure 29). Next, you are going to view a single gene, SLC25A3, to understand how the PGS Genome Browser can help you visualize differential expression and alternative splicing results.

- Type **SLC25A3** in the *Search* bar at the top of the window. The browser will browse to the gene (Figure 30)
- The *muscle*, *brain*, *heart*, *liver*, and genomic label tracks were described earlier in the tutorial. Here, the focus is on the *Isoform Proportion* track to explain how those color-coded tracks help to visualize differential expression and alternative splicing. The reads that are mapped to a certain tissue and the proportion of the transcript expressed in this tissue are colored the same. In this screenshot, *brain* is colored green, *heart* is colored yellow, *liver* is colored orange, and *muscle* is colored red

Figure 30: Viewing one gene (SLC25A3) in the Genome Viewer

SLC25A3 was reported by Wang, et al., (Nature, 2008) to have "mutually exclusive exons (MXEs)". The reads mapped to the transcripts of this gene in each of the tissue samples are shown in the browser as different tracks. The relative abundances of the individual transcripts of this gene are shown by the height of the color coded tracks. Note the transcript NM_213611 has low expression while transcripts NM_005888 and NM_002635 have higher expression. The alternative splicing pattern is shown, as the paper explained, indicating that different forms of the third exon are used in a tissue-specific manner.

Using the Track panel

At this point, you may find it useful to start altering track properties. Each track can be individually configured to alter the visualization of the reads. For example:

- Select a track and drag it to change the position of the track in the viewer
- Select the *brain* track, and a configuration panel will appear at the bottom of the track panel enabling the configuration of the sequence reads display, the Y-axis, the color, and the labels of the track
- Select a track and increase its height

Step 5 – Create a Gene List

To create a list of transcripts that are both significantly differentially expressed AND alternatively-spliced among the four tissue samples, use the **Create gene list** function from the workflow to invoke the **List Manager** (Figure 31). Each of the tabs (*Venn Diagram*, *ANOVA Streamlined*, and *Advanced*) can be used to generate combinations of lists.

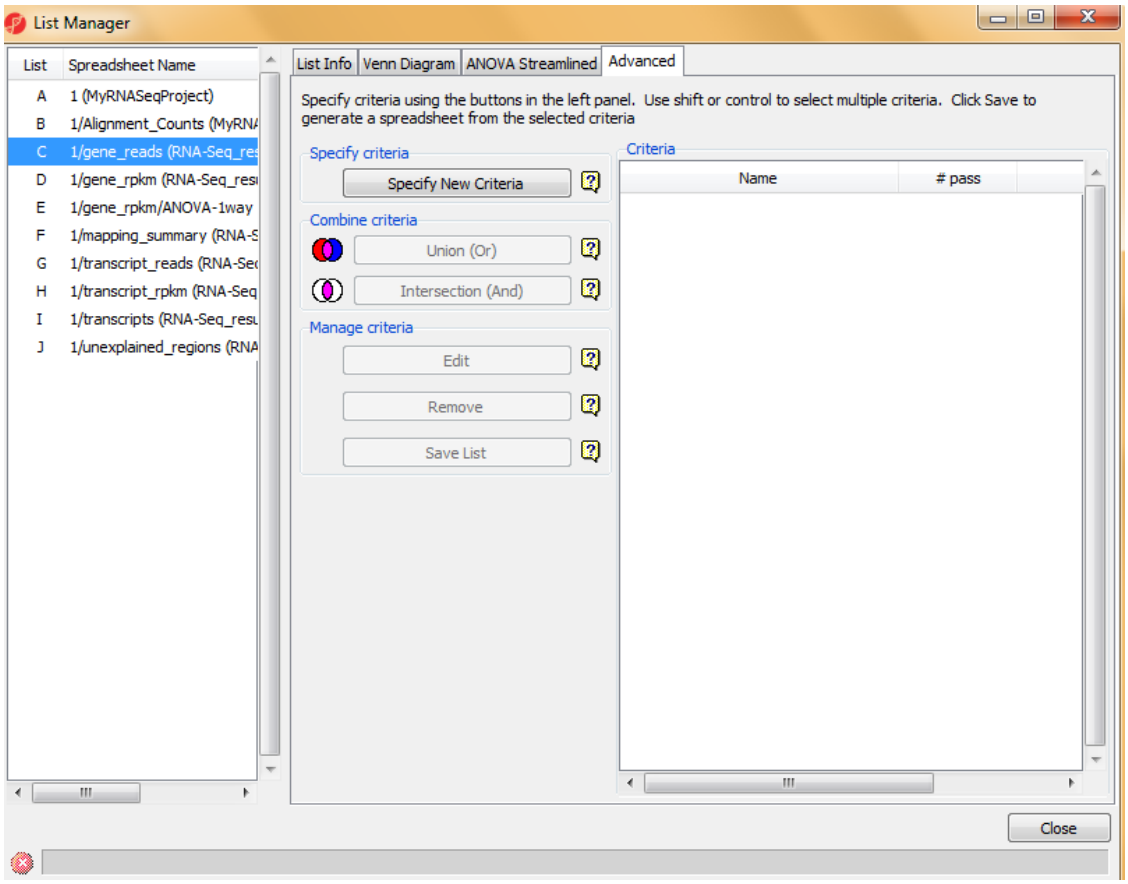


Figure 31: The List Manager

- Select the **Advanced** tab
- Select the **Specify New Criteria** button
- In the *Configure Criteria* dialog box (Figure 32), provide a name for the list (e.g., **Diff Exp**)
- Select the *transcripts* spreadsheet and the *p-value(DiffExp)* column
- Set *Include p-values* **significant with FDR** of **0.05**
- A list of 25,632 transcripts that pass this criteria will be generated
- Select **OK**

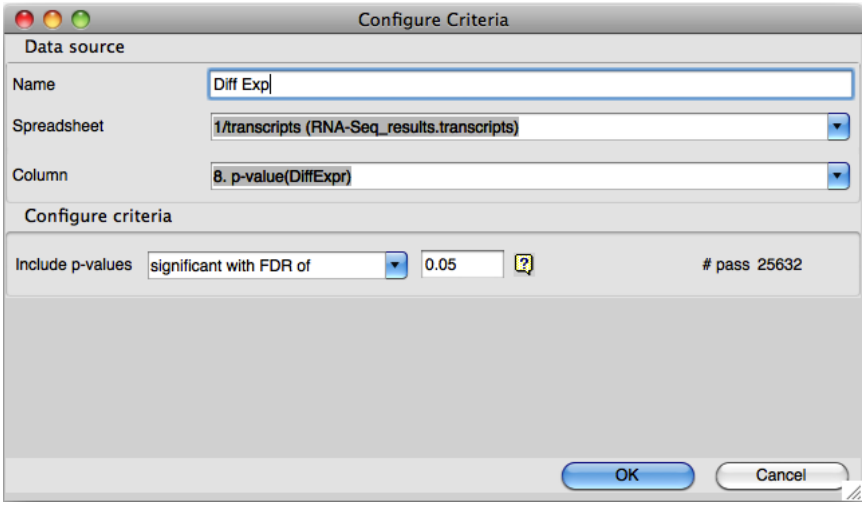


Figure 32: Configure Criteria dialog box

- These same steps can be used to create a list of transcripts likely alternatively spliced using the same FDR cutoff and the AltSplice p-value column
- Select both lists in the right panel under *Criteria* and then select **Intersection** from the left pane of the *List Manager*. Select **OK**. A list of 12,369 genes will be generated that includes all the genes that are both significantly differentially expressed and alternatively spliced among the four tissue samples
- Under *Manage Criteria*, select **Save List**
- Check the box for the intersection sheet of *Diff exp* and *Alt splice* and select **OK**. This list will now be available when you **Close** the *List Manager* (Figure 33)

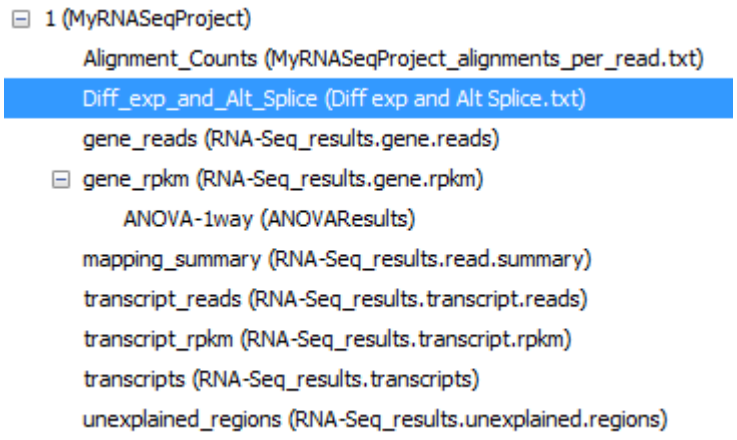


Figure 33: A list of the differentially expressed and alternatively spliced genes is now available for downstream analysis

The list of differentially expressed and alternatively spliced transcripts will be used in the next step, *Biological Interpretation*, for *GO Enrichment Analysis*.

Step 6 - Biological Interpretation: GO Enrichment

With the *GO Enrichment* feature in PGS, you can take a list of significantly expressed genes/transcripts and find GO terms that are significantly enriched within the list. For a detailed introduction of GO Enrichment, refer to the GO Enrichment User Guide (*Help > On-line Tutorials > User Guides*).

- Select **Gene set analysis** in the *Biological Interpretation* section of the workflow
- Select **GO Enrichment** and select **Next**. Choose the spreadsheet that was just created (Figure 34) and select Next

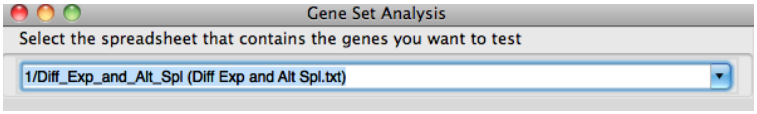

Gene Set Analysis
Select the spreadsheet that contains the genes you want to test
1/Diff_Exp_and_Alt_Spl (Diff Exp and Alt Spl.txt)

Figure 34: Selecting the spreadsheet for GO Enrichment

- The *Gene Set Analysis* dialog shown in Figure 35 allows you to select either *Fisher's Exact test* or the *Chi-Square test*. The *Chi-Square test* is faster than the *Fisher's Exact test* but is less accurate for sparse data. The defaults for the rest of the options are acceptable. Be sure to check **Invoke gene ontology browser on the result**. Select **Next**

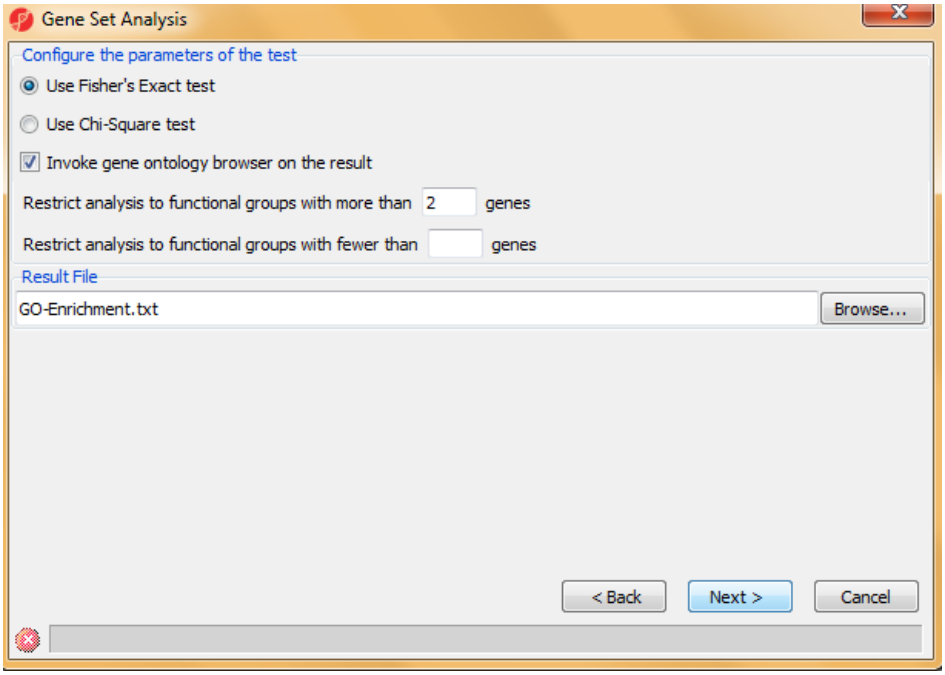

Gene Set Analysis
Configure the parameters of the test
Use Fisher's Exact test
Use Chi-Square test
Invoke gene ontology browser on the result
Restrict analysis to functional groups with more than 2 genes
Restrict analysis to functional groups with fewer than genes
Result File
GO-Enrichment.txt
Browse...
< Back
Next >
Cancel

Figure 35: GO Enrichment options

- Select **Default mapping file** and then select **Next**
- A *GO-Enrichment* spreadsheet, as well as a browser (Figure 36), will be generated with the enrichment score shown for each GO term
- Browse through the results to find a functional group of interest by examining the enrichment scores. The higher the enrichment score, the more overrepresented this functional group is in the input gene list. Alternatively, you may use the

Interactive filter on the *GO-Enrichment* spreadsheet to identify functional groups that have low p-values and perhaps a higher percentage of genes in the group that are present

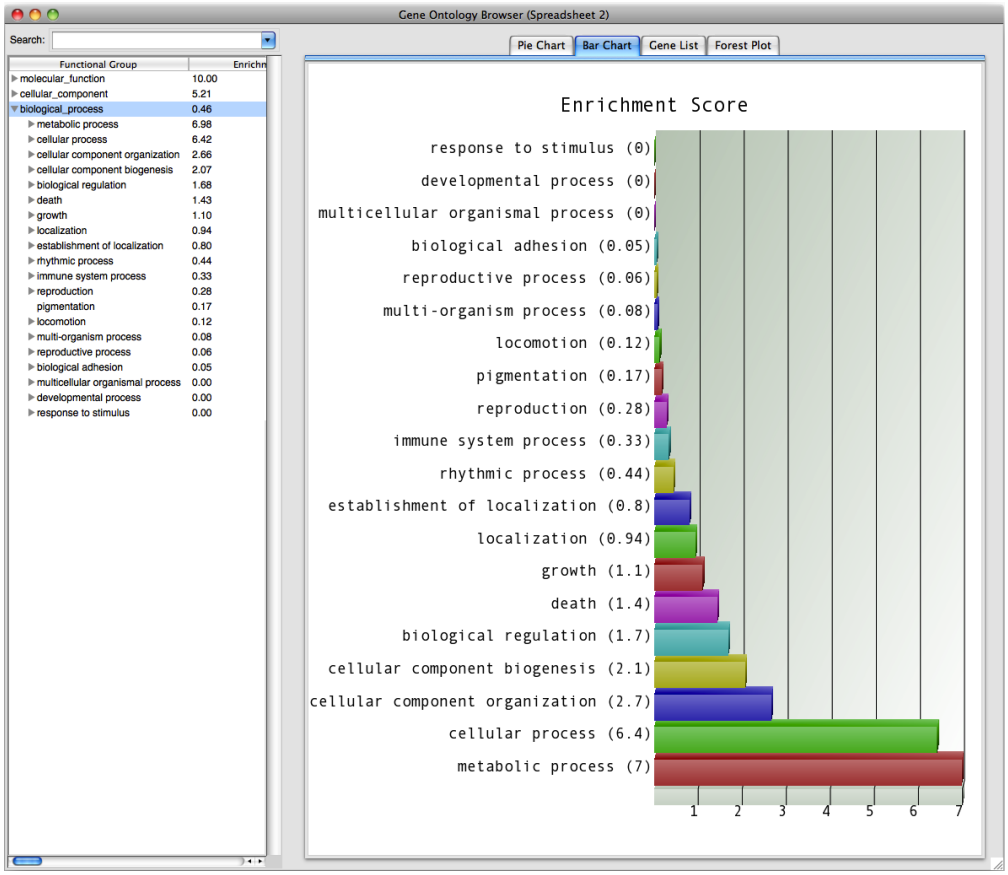


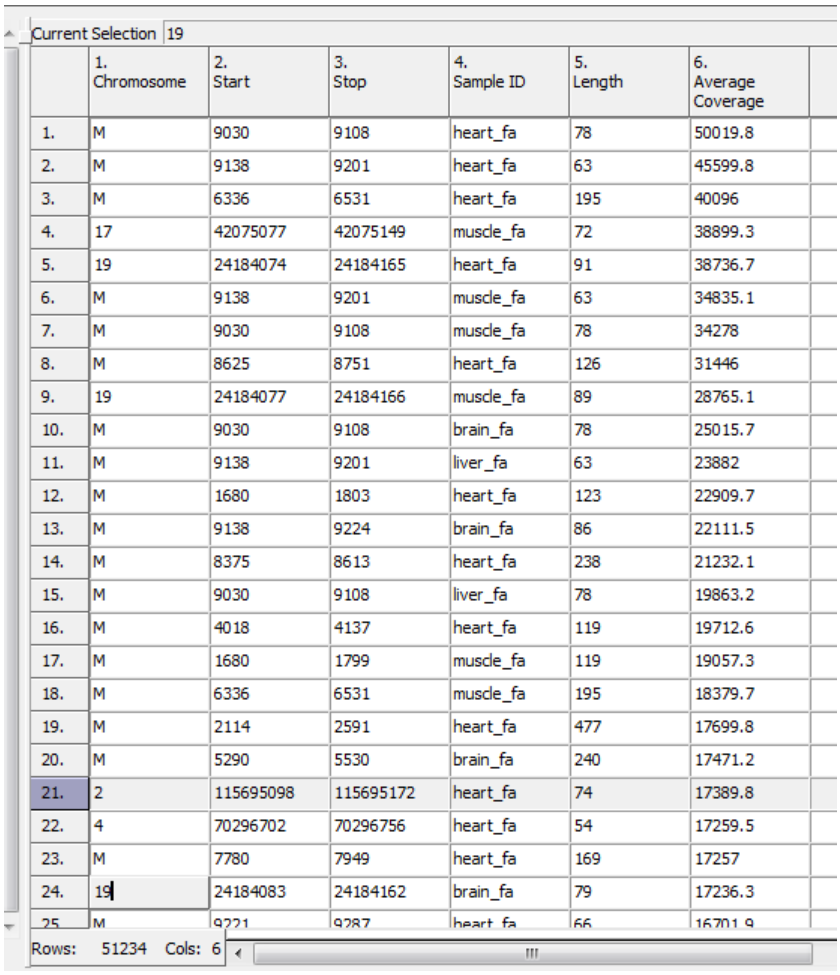
Figure 36: Viewing the Gene Ontology Browser

Step 7 - Detect Single Nucleotide Variations

To learn how to detect nucleotide differences and annotate SNP calls with overlapping genes, refer to the Analysis of Nucleotide Variations in NGS Data with Partek GS 6.6 tutorial using *Help > On-line Tutorials > Next-Gen Sequencing*.

Step 8 - Detecting Unexplained Regions

During **Step 3**, a spreadsheet named *unexplained_regions* was generated (Figure 37). This spreadsheet contains locations where reads map to the genome but are not annotated by the transcript database (RefSeqGene, in this case). The spreadsheet can be sorted by descending *Average Coverage* (column 6). This sheet is potentially very interesting as it may contain novel findings.



Current Selection 19

	1. Chromosome	2. Start	3. Stop	4. Sample ID	5. Length	6. Average Coverage
1.	M	9030	9108	heart_fa	78	50019.8
2.	M	9138	9201	heart_fa	63	45599.8
3.	M	6336	6531	heart_fa	195	40096
4.	17	42075077	42075149	muscle_fa	72	38899.3
5.	19	24184074	24184165	heart_fa	91	38736.7
6.	M	9138	9201	muscle_fa	63	34835.1
7.	M	9030	9108	muscle_fa	78	34278
8.	M	8625	8751	heart_fa	126	31446
9.	19	24184077	24184166	muscle_fa	89	28765.1
10.	M	9030	9108	brain_fa	78	25015.7
11.	M	9138	9201	liver_fa	63	23882
12.	M	1680	1803	heart_fa	123	22909.7
13.	M	9138	9224	brain_fa	86	22111.5
14.	M	8375	8613	heart_fa	238	21232.1
15.	M	9030	9108	liver_fa	78	19863.2
16.	M	4018	4137	heart_fa	119	19712.6
17.	M	1680	1799	muscle_fa	119	19057.3
18.	M	6336	6531	muscle_fa	195	18379.7
19.	M	2114	2591	heart_fa	477	17699.8
20.	M	5290	5530	brain_fa	240	17471.2
21.	2	115695098	115695172	heart_fa	74	17389.8
22.	4	70296702	70296756	heart_fa	54	17259.5
23.	M	7780	7949	heart_fa	169	17257
24.	19	24184083	24184162	brain_fa	79	17236.3
25.	M	9221	9287	heart_fa	66	16701.9

Rows: 51234 Cols: 6

Figure 37: The unexplained_regions spreadsheet

PGS can look for genes that overlap the regions. PGS locates reads not only outside of a known gene, it also takes exon/exon information from the database and locate reads to the intron of a gene. This is helpful for finding potential novel transcripts, exons, sequencing biases, etc.

- Go to **Tools > Find Overlapping Genes** option in the command toolbar
- In the resulting dialog box (Figure 38), select **Add a new column with the gene nearest to the region** and then **OK**

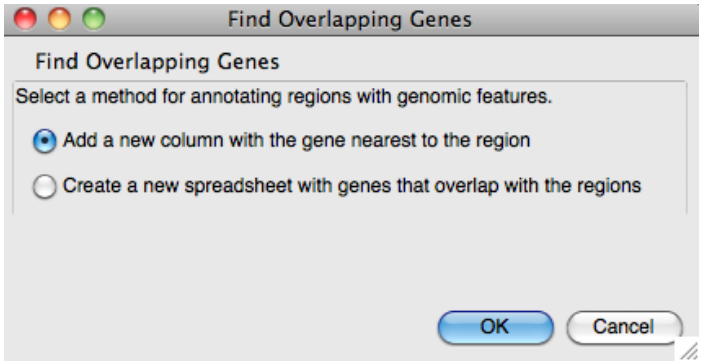
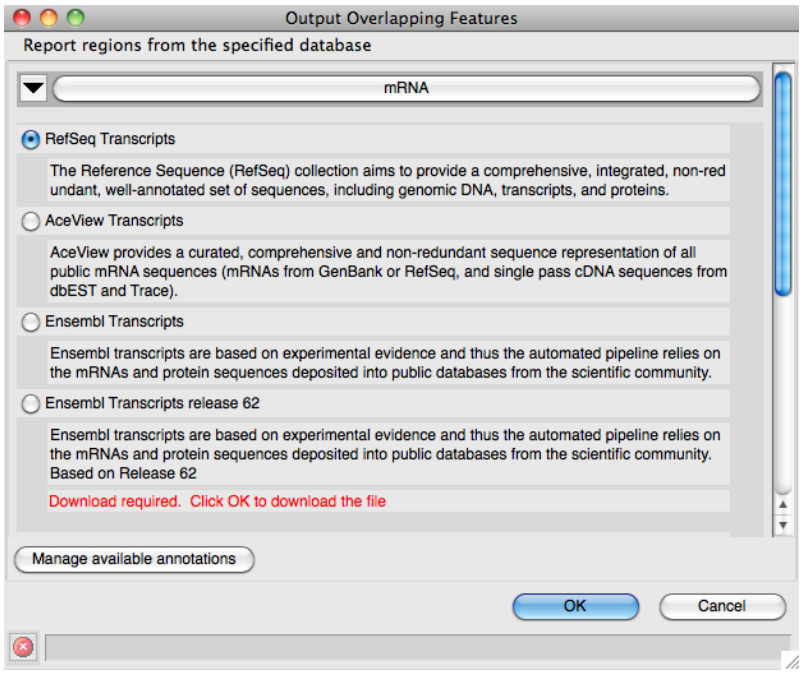

Find Overlapping Genes
Find Overlapping Genes
Select a method for annotating regions with genomic features.
Add a new column with the gene nearest to the region
Create a new spreadsheet with genes that overlap with the regions
OK
Cancel

Figure 38: Find Overlapping Genes

- Specify the database you wish to use. In this example, select **RefSeq Transcripts** (Figure 39)
- Select **OK**


Output Overlapping Features
Report regions from the specified database
mRNA
RefSeq Transcripts
The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.
AceView Transcripts
AceView provides a curated, comprehensive and non-redundant sequence representation of all public mRNA sequences (mRNAs from GenBank or RefSeq, and single pass cDNA sequences from dbEST and Trace).
Ensembl Transcripts
Ensembl transcripts are based on experimental evidence and thus the automated pipeline relies on the mRNAs and protein sequences deposited into public databases from the scientific community.
Ensembl Transcripts release 62
Ensembl transcripts are based on experimental evidence and thus the automated pipeline relies on the mRNAs and protein sequences deposited into public databases from the scientific community. Based on Release 62
Download required. Click OK to download the file
Manage available annotations
OK
Cancel
Figure 39: Select the database to search for overlapping features

The closest overlapping feature and the distance to it is now included (Figure 40) in the *unexplained_regions* spreadsheet.

Current Selection M

	1. Chromosome	2. Start	3. Stop	4. Sample ID	5. Length	6. Average Coverage	7. Overlapping Features	8. Nearest Feature	9. Distance to Nearest Feature (bps)
1.	M	9030	9108	heart_fa	78	50019.8	None	None	0
2.	M	9138	9201	heart_fa	63	45599.8	None	None	0
3.	M	6336	6531	heart_fa	195	40096	None	None	0
4.	17	42075077	42075149	muscle_fa	72	38899.3	intron of PYY	PYY (-)	0
5.	19	24184074	24184165	heart_fa	91	38736.7	region starts	ZNF726 (+)	77556
6.	M	9138	9201	muscle_fa	63	34835.1	None	None	0
7.	M	9030	9108	muscle_fa	78	34278	None	None	0
8.	M	8625	8751	heart_fa	126	31446	None	None	0
9.	19	24184077	24184166	muscle_fa	89	28765.1	region starts	ZNF726 (+)	77559
10.	M	9030	9108	brain_fa	78	25015.7	None	None	0
11.	M	9138	9201	liver_fa	63	23882	None	None	0
12.	M	1680	1803	heart_fa	123	22909.7	None	None	0
13.	M	9138	9224	brain_fa	86	22111.5	None	None	0
14.	M	8375	8613	heart_fa	238	21232.1	None	None	0
15.	M	9030	9108	liver_fa	78	19863.2	None	None	0
16.	M	4018	4137	heart_fa	119	19712.6	None	None	0
17.	M	1680	1799	muscle_fa	119	19057.3	None	None	0
18.	M	6336	6531	muscle_fa	195	18379.7	None	None	0
19.	M	2114	2591	heart_fa	477	17699.8	None	None	0
20.	M	5290	5530	brain_fa	240	17471.2	None	None	0
21.	2	115695098	115695172	heart_fa	74	17389.8	intron of	DPP10 (+)	0

Figure 40: The unexplained_regions spreadsheet showing regions mapped to the closest genomic features

The description of each column in the *unexplained peaks* spreadsheet is shown in Table 3.

Table 3: Description of annotated columns in unexplained_regions

Column Label	Description
Chromosome, Start, Stop	Genomic location of the region containing the reads
Sample ID	Sample that contains the reads mapped to this region
Length	Length of the region (in base pairs)
Average Coverage	Average read coverage in the region
Overlapping Features	Section of the nearest gene that overlaps with the region (intron, starts before or after a gene, region contained in a gene, gene contained within the region, region overlaps with a gene)
Nearest Feature	Name of the nearest feature and strand (+ or -)
Distance to Nearest Feature (bps)	Distance of the detected region to the closest gene. If the detected region is mapped to the intron of a gene, the distance is shown as 0

Right-clicking on a row header and selecting **Browse to Location** will show the reads mapped to the chromosome. For this tutorial, a couple of genes are selected to show regions that are located after a known gene or in the intron of a gene.

- With the *unexplained_regions* spreadsheet open, right-click on *Average Coverage* (column 6) and select **Sort Descending**
- Select row **45** and **Browse to location** to show a region within an intron of UNC45B (Figure 41, left panel). This may be a novel exon
- Select row **10744** and **Browse to location** to show a region that starts 1 bp after CD82 (Figure 41, right panel). This peak may represent an extended exon

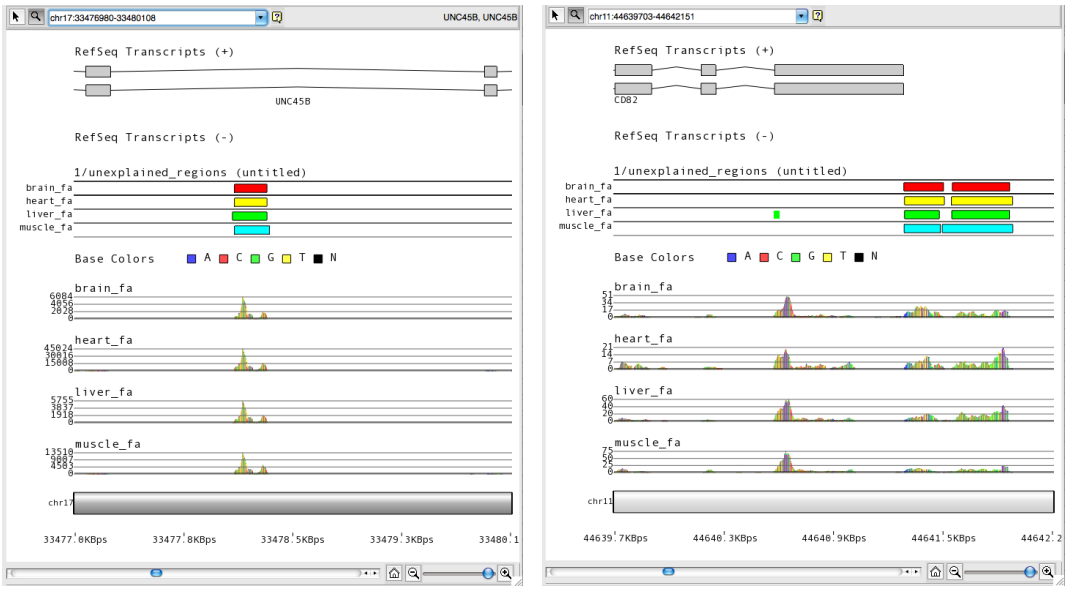


Figure 41: A region view of reads mapped to non-overlapping genes

While RefSeq was used to identify overlapping features, the choice of which database to use will depend on the biological context of your experiment. For example, you may wish to utilize promoter or microRNA databases if you are interested in regulation of expression.

End of Tutorial

This is the end of RNA-Seq tutorial. If you need additional assistance with this data set, contact the Partek Technical Support staff at +1-314-878-2329 or email us at support@partek.com.

Date last updated: Feb. 2012

Copyright © 2012 by Partek Incorporated. All Rights Reserved. Reproduction of this material without express written consent from Partek Incorporated is strictly prohibited.

References

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature, 2008; 5: 621-8.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., & Burge, C.B. Alternative isoform regulation in human tissue transcriptomes. Nature, 2008; 456: 470-6.

Xing Y, Yu T, Wu YN, Roy M, Kim J, Lee C: An expectation-maximization algorithm for probalisitic reconstructions of full-length isoforms from splice graphs. Nucleic Acids Res 2006, 34: 3150-3160.